Drug Information Journal, Vol. 34, pp. 605–615, 2000
Printed in the USA.

0092-8615/2000

PROPERTIES OF THE HUI AND WALTER AND RELATED METHODS FOR ESTIMATING PREVALENCE RATES AND ERROR RATES OF DIAGNOSTIC TESTING PROCEDURES

MICHAEL D. SINCLAIR, PHD
Senior Statistician, Mathematica Policy Research, Princeton, New Jersey
JOSEPH L. GASTWIRTH, PHD
Visiting Scientist, Division of Cancer Epidemiology and Genetics, National Cancer Institute, and Professor of Economics and Statistics, The George Washington University Department of Statistics, Washington, District of Columbia

When a confirmatory test is completely accurate or has known low error rates, the sensitivity and the specificity of a screening test can be estimated. When the error rates for the confirmatory test are unknown, Hui and Walter (2) presented a method for estimating the sensitivity and specificity of both the screening and the confirmatory tests using the tests on two populations with different prevalence rates of the infection. The method requires that the tests have equal error rates in the two populations. When this requirement is not met, we show that the estimated prevalence rates are robust when the difference in the prevalence rates of the two subpopulations is large. An alternative design, requiring only one population, but other assumptions, is also described.

Key Words: Diagnostic tests; Classification errors; Sensitivity; Specificity; Evaluation design

INTRODUCTION

A RECENT PAPER BY Hui and Xiao (1) discusses the evaluation of diagnostic tests when a gold standard or perfect confirmatory test is not available. Hui and Walter (2) developed a method to estimate the error rates in a diagnostic test when it and a second test are applied to two groups of individuals. This procedure requires two subpopulations with different prevalence rates in which each test has the same error rates. Goldberg (3) noted that obtaining two subpopulations satisfying this criteria may be difficult. For example, Gastwirth (4) cited data showing that the accuracy of the early ELISA screening test for the HIV virus differed in men and women.

This paper explores the bias in the Hui and Walter (2) method when the assumption of equal classification errors in the two subpopulations is violated. First, we will review the Hui and Walter (2) method and discuss its properties when the assumption of equal classification errors is violated. Second, we present an alternative design using only one population. Finally, we discuss a numerical study which examines the sensitivity of the

Reprint address: Dr. Michael D. Sinclair, Mathematica Policy Research, 600 Alexander Park, P.O. Box 2393, Princeton, NJ 08543–2393.

estimated accuracy rates to violations in the assumptions required for both methods.

Study Designs for Estimating the Accuracy of Screening Tests

We will assume that two tests with dichotomous responses are applied to a sample of n_g individuals from each of two subpopulations g, g = 1 or 2. The true classification status of each sample unit is unknown and each test has unknown sensitivity and specificity. Following Hui and Walter (2), we let π_g denote the true prevalence rate of the individuals in subpopulation g and let α_{rg} and β_{rg} denote the false positive and false negative rates, respectively, associated with each test r, r= 1 or 2, for subpopulation g. The respective specificities and sensitivities of the two tests are one minus each of the error rate parameters. For each subpopulation g, the sample frequencies associated with each of the four possible combinations of Test 1 and Test 2 results can be represented by a 2×2 table as in Figure 1. In this table we let P_{gij} denote the probability of obtaining a combination of Test 1 and Test 2 responses associated with row i and column j of the 2×2 table for subpopulation g, and similarly n_{gij} as the observed frequency of test outcomes in each cell for a given sample size n_g from subpopulation g.

With this notation and assuming conditional independence between the two test error rates, the probabilities, P_{gij}, are as in (1).

$$
\begin{aligned}
P_{g11} &= \pi_g (1-\beta_{1g})(1-\beta_{2g}) + (1-\pi_g)(\alpha_{1g} \cdot \alpha_{2g}) \\
P_{g21} &= \pi_g (\beta_{1g})(1-\beta_{2g}) + (1-\pi_g)(1-\alpha_{1g})(\alpha_{2g}) \\
P_{g12} &= \pi_g (1-\beta_{1g})\,\beta_{2g} + (1-\pi_g)(\alpha_{1g})(1-\alpha_{2g}) \\
P_{g22} &= \pi_g (\beta_{1g} \cdot \beta_{2g}) + (1-\pi_g)(1-\alpha_{1g})(1-\alpha_{2g}).
\end{aligned} \tag{1}
$$

Hui and Walter (2) assume that $\beta_{r1} = \beta_{r2}$, $\alpha_{r1} = \alpha_{r2}$, for r=1 and 2, and $\pi_1 \neq \pi_2$, that is, the test errors are equal for both subpopulations, but the prevalence rates differ. When two subpopulations satisfying these assumptions are available, the number of independent cell probabilities is equal to the number of parameters (six in total), so estimation is possible. Parameter estimates in terms of the observed cell probabilities, denoted by p_{gij}, are presented along with their variances obtained from the asymptotic information matrix in Hui and Walter's (2) paper.

A New Design Using One Population

In this section we present an alternate design which may allow for the evaluation of two testing procedures when two populations

Test 1 Outcome	Test 2 Outcome		
	Positive	Negative	Total
Positive	n_{g11}	n_{g12}	$n_{g1.}$
Negative	n_{g21}	n_{g22}	$n_{g2.}$
Total	$n_{g.1}$	$n_{g.2}$	$n_{g..}$

FIGURE 1. Cross classification of Test 1 and Test 2 results.

with equal test error rates are not available. As in the Hui and Walter (2) method, we assume that the test error rates are independent.

In this method a sample of n individuals is randomly divided into two groups. Since the focus of these procedures is to evaluate the screening test, we administer the screening test twice to each individual in group 1, and the screening test and a second test (preferably a confirmatory test) to each individual in Group 2. With this design, we have a common prevalence rate, $\pi = \pi_1 = \pi_2$, the error rates for the screening test, $\alpha_1 = \alpha_{11} = \alpha_{12} = \alpha_{21}$ and $\beta_1 = \beta_{11} = \beta_{12} = \beta_{21}$ and error rates for the second test applied to group 2, $\alpha_2 = \alpha_{22}$ and $\beta_2 = \beta_{22}$ to yield a total of five parameters. The two 2×2 tables associated with this design contain only a total of four degrees of freedom (There are still six degrees of freedom, but two are redundant, and so only four are available for estimation). Hence, in order to estimate the accuracy rates an additional assumption is needed. Therefore, we assume that the gain in both the specificity and sensitivity of the second test relative to the first to be equal. In other words, the second test in the second group either lowers or increases the false positive and false negative rate associated with the screening test by an unknown factor, c, such that, $\beta_2 = c \times \beta_1$, and $\alpha_2 = c \times \alpha_1$ for $c > 0$. We will refer to this condition as the equal fraction reduction requirement (EFR).The expected probabilities associated with this model are given in (2)

$$\begin{aligned}
P_{g11} &= \pi\,[(1-\beta_1)(1-\beta_g)] + (1-\pi)[(\alpha_1\,\alpha_g)] \\
&= \pi\,[(1-\beta_1)(1-c_g\beta_1)] + (1-\pi)\,[(c_g\alpha_1^2)] \\
P_{g12} &= \pi\,[(\beta_1)(1-\beta_g)] + (1-\pi)[(1-\alpha_1)(\alpha_g)] \\
&= \pi\,[(\beta_1)(1-c_g\beta_1)] + (1-\pi)[(1-\alpha_1)\,(c_g\alpha_1)] \\
P_{g21} &= \pi\,[(1-\beta_1)\beta_g] + (1-\pi)[(\alpha_1)(1-\alpha_g)] \\
&= \pi\,[(1-\beta_1)c_g\beta_1] + (1-\pi)[(\alpha_1)\,(1-c_g\alpha_1)] \\
P_{g22} &= \pi\,[(\beta_1\,\beta_g)] + (1-\pi)[(1-\alpha_1)(1-\alpha_g)] \\
&= \pi\,[(c_g\beta_1^2)] + (1-\pi)[(1-\alpha_1)\,(1-c_g\alpha_1)] \qquad (2)
\end{aligned}$$

for g = 1 and 2, where $c_g = 1$ for subpopulation g = 1 given a repeat of the screening test. As with the Hui and Walter (2) model, we now have a saturated model for which unique parameter estimation can be conducted. Closed form maximum likelihood estimates for the model parameters are not obtainable so estimates are obtained by numerical techniques.

METHODS

We conducted a numerical study of the bias in the Hui and Walter (2) method when the two subpopulations selected do not have equal classification errors and a similar study for the proposed single population design when the second test, preferably a confirmatory test, does not reduce the false positive and false negative rates in the same proportion. We used two approaches to examine the robustness of each design when the required error rate assumptions were violated. First, we generated the expected cell frequencies that would be obtained in each of two subgroups for a specified sample size of 1000 (or 4000 for the smaller prevalence rates studied) individuals in each subgroup. We generated the expected cell frequencies based on a variety of specified parameter values using the cell probabilities given in equations (1) or (2). To keep the number of possible parameter values to a manageable level, we limited the parameter values for the prevalence rate in each subgroup to paired values of .20 and .40 or .02 and .04 and examined error rates between .01 and .10. Second, we conducted a Monte-Carlo simulation in which we randomly generated combined test outcomes for 1000 (or 4000) individuals in each subgroup based on the same parameter values. We repeated this process for 2000 (or 4000 for the smaller prevalence rates) iterations to generate 2000 2×2 tables with 1000 observations

in each (or 4000 2×2 tables with 4000 observations in each) for each of the two subgroups. At each iteration, we prepared estimates of the parameters appropriate for each design from the pair of tables. Ultimately, this produced 2000 (or 4000) sets of estimates for each set of parameter values evaluated for each design. From these sets, we obtained the mean value and a standard error of the parameter estimates.

Because the sampling variability in the Monte-Carlo method is large relative to the bias in the estimates, we use the parameter estimates obtained from the expected cell frequencies to estimate the bias in the estimates and the standard errors obtained from the Monte-Carlo simulation to estimate the statistical precision in the estimates. The Monte-Carlo estimates of the standard error in the estimates are needed because the corresponding estimates from the expected frequency approach are based on the design assumptions which are violated in the samples to address the robustness of the procedures. As indicated above, the iterations associated with the smaller prevalence rates are more variable, so we conducted 4000 rather than 2000 iterations, and used a sample size of 4000 rather than 2000 to help stabilize the results.

We prepared the parameter estimates in each situation using the SAS (5) NLIN Gauss-Newton weighted least squares procedure. A general description of this method can be found in Bard (6) and Jennrich and Moore (7). Bradley (8) showed these procedures to be equivalent to maximum likelihood estimates for the class of probability distributions including equation (1) and (2). Further details on the estimation methods are discussed in Sinclair (9).

RESULTS

Numerical Estimates of the Bias in the Hui and Walter Estimates When the Two Subpopulations Do Not Have the Same Error Rates

The bias in the estimates obtained from the Hui and Walter (2) method are given in Tables 1 and 2 and Figures 2 and 3. In this study, the false positive and false negative rates from Test 1 for the first subpopulation were set to 5% ($\alpha_1 = \beta_1 = .05$) and these values for Test 2, a confirmatory test, were set to 2% ($\alpha_2 = \beta_2 = .02$). Figures 2 and 3 present the bias in the estimated prevalence rate for Population 1 when the screening test error rates in the second population differ from those in the first by less than or equal to 0.05 (less than a unit multiple of the first population values). In Figures 2 and 3, we set the confirmatory test error rate to be the same in both subpopulations so that only the screening test error rates violated the Hui and Walter (2) assumptions. While these graphs are specific to these examples, they illustrate the general bias properties of the Hui and Walter (2) method for the prevalence rate in the presence of a violation of the equal error rate assumption.

A study of both the bias in the estimated prevalence rate and error rates is presented in Tables 1 and 2. Eight numerical examples are presented in each table. In examples one through four, as in Figures 2 and 3, only the screening test error rates are larger or smaller in the second subpopulation and the error rates of the confirmatory test remain the same. In examples five through eight, both the screening test and the confirmatory test error rates are larger or smaller in the second subpopulation. Columns (2) and (3) of Tables 1 and 2 present the estimated prevalence rates from the screening and confirmatory tests, respectively. These estimates do not account for the error rates of the tests and are clearly biased. Note that Figures 2 and 3 explore the bias in the prevalence rate for the first population for a broader range of combinations than those represented by examples 1 to 4 in Tables 1 and 2.

Table 1 considers prevalence rates of .20 and .40. Example A illustrates the unbiasedness of the Hui and Walter estimates when the error rates are the same for both subpopulations. By comparing examples one and two (five and six) to examples three and four (seven and eight), we notice that the bias in the estimates is more severe when the larger

TABLE 1
Bias in Hui and Walter Estimates When the Error Rates in the Two Subpopulations Selected Differ Using Prevalence Rates of 20% and 40% (Sample Size = 1000 in Each Subpopulation)

EXAMPLE	True Values			Screening Test Estimates	Uncorrected Confirmatory Test Estimates			Hui and Walter Estimates Standard Error of the Estimates in ( )s		
	Error Rates		Prevalence Rates	Prevalence Rates	Prevalence Rates	Screening Error Rates		Prevalence Rates	Error Rates (Assumed Equal Population 1 & Population 2)	Bias in Prevalence
	Population 1	Population 2				Population 1	Population 2			
	α_1 α_2 β_1 β_2	α_1 α_2 β_1 β_2	π_1 π_2	π_1 π_2	π_1 π_2	α_1 β_1	α_1 β_1	π_1 π_2	α_1 α_2 β_1 β_2	π_1 π_2
A	.05 .02	.05 .02	.20	.230	.212	.05457	.06208	.20000 (.01585)	.05000 (.01178) .02000 (.01122)	.00000
	.05 .02	.05 .02	.40	.410	.404	.11792	.07673	.40000 (.02050)	.05000 (.02494) .02000 (.02511)	.00000
1	.05 .02	.060 .02	.20	.230	.212	.05457	.07181	.20906 (.01768)	.04389 (.01376) .01566 (.01109)	.00906
	.05 .02	.060 .02	.40	.412	.404	.11792	.08614	.41350 (.02172)	.06589 (.00249) .04521 (.03065)	.01350
2	.05 .02	.075 .02	.20	.230	.212	.05457	.08641	.22329 (.01846)	.03432 (.01511) .00891 (.01023)	.02329
	.05 .02	.075 .02	.40	.415	.404	.11792	.10025	.43439 (.02201)	.08932 (.02321) .08156 (.03207)	.03439
3	.05 .02	.040 .02	.20	.230	.212	.05457	.05235	.19251 (.01461)	.05345 (.00885) .02414 (.01074)	−.00749
	.05 .02	.040 .02	.40	.407	.404	.11792	.06733	.38926 (.01852)	.03403 (.02313) .00000 (.01613)	−.01074
4	.05 .02	.025 .02	.20	.230	.212	.05457	.03775	.18789 (.01327)	.04613 (.00599) .02969 (.00867)	−.01211
	.05 .02	.025 .02	.40	.405	.404	.11792	.05322	.38576 (.01608)	.01064 (.01708) .00000 (.00463)	−.01424
5	.05 .02	.060 .024	.20	.230	.212	.05457	.07419	.21247 (.01764)	.04237 (.01431) .01323 (.01094)	.01247
	.05 .02	.060 .024	.40	.412	.405	.11792	.09130	.41856 (.02171)	.07452 (.02449) .05125 (.03113)	.01856
6	.05 .02	.075 .03	.20	.230	.212	.05457	.09217	.23207 (.01820)	.03041 (.01521) .00270 (.00887)	.03207
	.05 .02	.075 .03	.40	.415	.406	.11792	.11268	.44717 (.02122)	.10953 (.02130) .09540 (.03190)	.04717
7	.05 .02	.040 .016	.20	.230	.212	.05457	.04987	.19062 (.01359)	.05221 (.00834) .02642 (.01034)	−.00938
	.05 .02	.040 .016	.40	.407	.403	.11792	.06190	.38701 (.01798)	.02484 (.02172) .00000 (.01378)	−.01299
8	.05 .02	.025 .010	.20	.230	.212	.05457	.03136	.18597 (.01253)	.04299 (.00556) .02942 (.00600)	−.01403
	.05 .02	.025 .010	.40	.405	.402	.11792	.03918	.38547 (.01591)	.00000 (.00909) .00000 (.00149)	−.01453

of the error rates is associated with the larger of the two prevalence rates. In example one, the screening test error rates in the second subpopulation (with a prevalence rate of .40) are larger (α_1 and $\beta_1 = .05$ in pop. one, and α_1 and $\beta_1 = .06$ in pop. two), while in example 3 the error rates in the second population are smaller (α_1 and $\beta_1 = .05$ in pop. one, and α_1 and $\beta_1 = .04$ in pop. two). The Hui and Walter estimates of the prevalence rate in example one are .20906 and .41350 with a bias of .00906 and .01350, respectively. In Example 3, the estimates of the prevalence rate are .19251 and .38926 with a bias of −.00749 and −.01074. This is even more apparent in Figure 2. The bias in the quadrant in which the second population error rates are higher is much larger than the bias in the quadrant associated with smaller error rates for Population 2.

Assuming that the screening test error rates are higher than those of the confirmatory test, a large portion of the bias in the estimates of the prevalence rate in these examples is due to the differences in the error rates of the screening tests in the two subpopulations. In Example 5, when both the error rates are increased, the estimates of the prevalence rate are .21247 and .41856 with a respective bias of .01247 and .01856. Comparing this with the results in Example 1 (.00906 and .01350), the level of bias is only modestly larger.

Table 2 presents examples for prevalence rates of .02 and .04. (Note that in these tables the results are based on a sample size of 4000 individuals from each subpopulation). As in Table 1, the bias in the estimates is more severe when the larger of the error rates is associated with the larger of the two prevalence rates. The effect is more substantial, however, than that examined in Table 1. In example 1, the larger error rates are associated with a prevalence rate of .04, which yields estimates of the prevalence rate at .03123 and .06178 with a bias of .01123 and .02178, respectively. In contrast, in example three, in which the larger error rates are associated with the prevalence rate of .02, we obtain prevalence rate estimates of .01949 and .03908 with bias levels of −.00051 and −.00092. This relationship between the size of the error rates and the prevalence rates can be seen by comparing Figures 2 and 3. Overall, as the difference in the prevalence rates in the two subpopulations increases, the bias in the estimates becomes smaller. Using prevalence rates of .20 and .60 (not presented), we found that the bias in the prevalence rate estimates becomes quite small and that the relationship between the size of the error rates and the prevalence rates on the bias of the prevalence rate estimates is also smaller.

In contrast with the prevalence rates estimates obtained from the Hui and Walter (2) method, the estimated accuracy of the screening and confirmatory tests appear to be sensitive to the violation in the equal error rate assumption. Although the estimated sensitivity and specificity of the screening test is noticeably more robust than those for the confirmatory test, they still appear to be rather sensitive. In Example 1 in Table 1, the screening test error rates are 20% higher in subpopulation 2 than in subpopulation 1 (α_1 and $\beta_1 = .05$ in Population 1 and α_1 and $\beta_1 = .06$ in Population 2), to yield estimated screening test error rates of .04389 and .06589 for α_1 and β_1, respectively. Hence a fairly modest violation in the required assumptions created a relatively large difference between the estimated and actual values. Furthermore, the estimated error rates for each test do not lie between the actual values in the two subpopulations.

As indicated, the confirmatory test error rate estimates are considerably more sensitive. As the level of the violation increases, one of the confirmatory test error rates will approach or lie on the zero boundary while the other will become larger than .05. In example 3, in Table 1, the screening test error rates are smaller in the second subpopulation (α_1 and $\beta_1 = .05$ in Population 1 and α_1 and $\beta_1 = .04$ in Population 2) and the confirmatory error rates remain the same in both subpopulations at .02. For this example, the Hui and

TABLE 2
Bias in Hui and Walter Estimates When the Error Rates in the Two Subpopulations Selected Differ Using Prevalence Rates of 2% and 4% (Sample Size = 2000 in Each Subpopulation)

EXAMPLE	True Values			Screening Test Estimates	Uncorrected Confirmatory Test Estimates			Hui and Walter Estimates		
	Error Rates		Prevalence Rates	Prevalence Rates	Prevalence Rates	Screening Error Rates		Prevalence Rates	Error Rates (Assumed Equal Population 1 & Population 2)	Bias
	Population 1	Population 2				Population 1	Population 2			
	α_1 α_2 β_1 β_2	α_1 α_2 β_1 β_2	π_1 π_2	π_1 π_2	π_1 π_2	α_1 β_1	α_1 β_1	π_1 π_2	α_1 α_2 β_1 β_2	π_1 π_2
A	.05 .02	.05 .02	.02	.068	.039	.05038	.05077	.02000 (.00611)	.05000 (.00541) .02000 (.00399)	.00000
	.05 .02	.05 .02	.04	.086	.058	.50000	.34589	.04000 (.00919)	.05000 (.09091) .02000 (.11423)	.00000
1	.05 .02	.060 .02	.02	.068	.039	.05038	.06075	.03123 (.01806)	.04019 (.01492) .01957 (.00827)	.01123
	.05 .02	.060 .02	.04	.095	.058	.50000	.34932	.06178 (.02014)	.06939 (.09503) .35190 (.13139)	.02178
2	.05 .02	.075 .02	.02	.068	.039	.05038	.07572	.04938 (.07125)	.02466 (.05653) .01890 (.02734)	.02938
	.05 .02	.075 .02	.04	.109	.058	.50000	.35445	.09610 (.07126)	.09766 (.12862) .57008 (.08261)	.05610
3	.05 .02	.040 .02	.02	.068	.039	.05038	.04078	.01949 (.00366)	.04494 (.00246) .02010 (.00376)	−.00051
	.05 .02	.040 .02	.04	.077	.058	.50000	.34247	.03908 (.00518)	.03962 (.08598) .00000 (.01666)	−.00092
4	.05 .02	.025 .02	.02	.068	.039	.05038	.02510	.01935 (.00350)	.03646 (.00215) .02024 (.00366)	−.00065
	.05 .02	.025 .02	.04	.063	.058	.50000	.33733	.03895 (.00504)	.02460 (.08367) .00000 (.00000)	−.00105
5	.05 .02	.060 .024	.02	.068	.039	.05038	.06090	.03728 (.02335)	.04010 (.01398) .01571 (.00984)	.01728
	.05 .02	.060 .024	.04	.095	.062	.50000	.38660	.07360 (.02825)	.21099 (.11413) .35430 (.13093)	.03360
6	.05 .02	.075 .03	.02	.068	.039	.05038	.07103	.07042 (.15285)	.02497 (.10117) .00939 (.06669)	.03861
	.05 .02	.075 .03	.04	.109	.068	.50000	.42153	.13751 (.15180)	.36391 (.16079) .56730 (.11717)	.07473
7	.05 .02	.040 .016	.02	.068	.039	.05038	.04062	.01876 (.00248)	.04484 (.00249) .01912 (.00209)	−.00124
	.05 .02	.040 .016	.04	.077	.055	.50000	.29825	.03757 (.00357)	.00000 (.03754) .00000 (.01516)	−.00243
8	.05 .02	.025 .010	.02	.068	.039	.05038	.02540	.01906 (.00215)	.03617 (.00218) .01531 (.00145)	−.00094
	.05 .02	.025 .010	.04	.063	.049	.50000	.21037	.03832 (.00308)	.00000 (.00153) .00000 (.00000)	−.00168

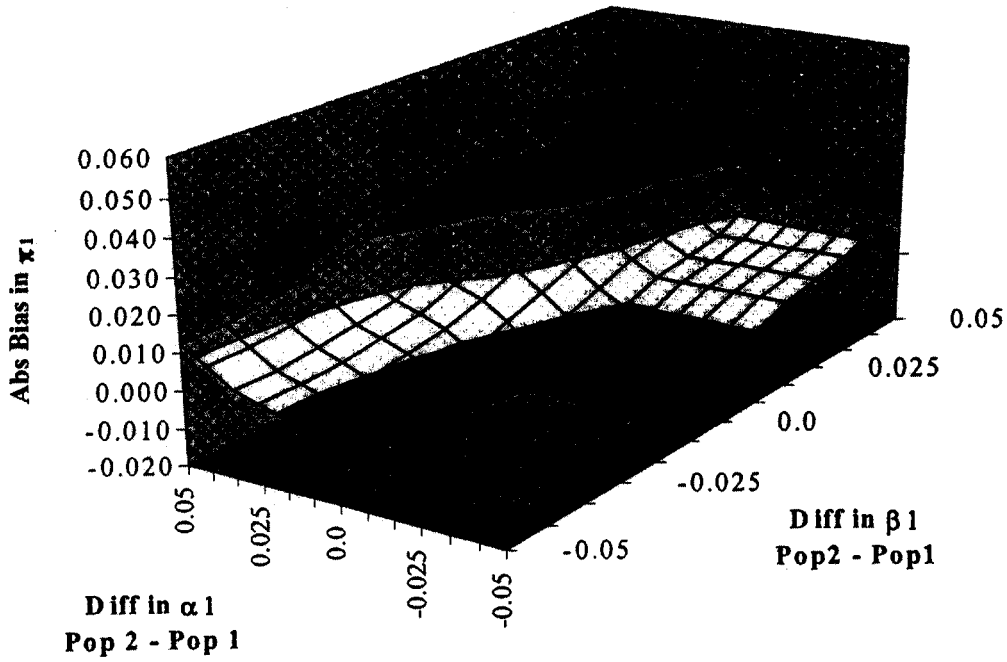

FIGURE 2. Bias in the first subpopulation estimate of prevalence when error rates in the two subpopulations differ. Actual prevalence in the first population is .20 (20%).

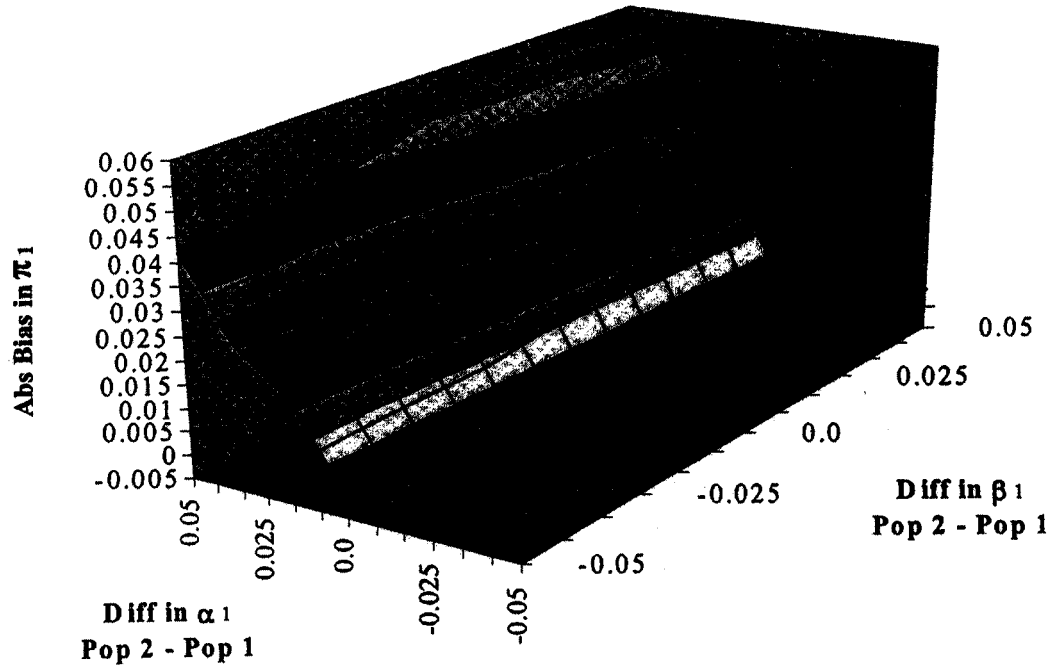

FIGURE 3. Bias in the first subpopulation estimate of prevalence when error rates in the two subpopulations differ. Actual prevalence in the first subpopulation is .02 (2%).

Walter (2) estimates of α_1, β_1, α_2, and β_2 are .05345, .03403, .02414, and .00000, respectively.

Numerical Estimates of the Bias in the Single Population Method

Table 3 presents the bias in the estimates obtained from the single population design when the false positive and the false negative rates for the confirmatory test are not an equal fraction of the screening test error rates. Ten examples are presented for a screening test that has a false positive rate and a false negative rate of .05. In Example 1, the confirmatory test in the second subpopulation reduces both the screening test's false positive and false negative rate to one fourth (25%) of the screening test rate. Hence, the single population method's EFR assumption is met and the estimates are unbiased (note the single study method estimate of c is equal to .25). In the remaining examples the EFR assumption is violated by various degrees. As for the Hui and Walter (2) method, we have prepared some examples using the underlying prevalence rate of .20 (Examples 2 to 7) and .02 (Examples 8 to 10).

Overall, the estimated false positive rate (1 – specificity) of the screening test (α_i) does not appear to be sensitive to a modest violation in the EFR assumption. The estimated false negative rate (1 – sensitivity) for the screening test (β_i) is somewhat more sensitive in these examples, but still yields estimates reasonably close to .05. Even in examples 4, 5, 9, and 10, in which the EFR assumption is violated by a factor of 2, the bias in this estimate is less then .022. Of course, in low prevalence situations (Examples 8 to 10) where there are relatively few positive cases, the standard error of the estimated false negative rate is large. Furthermore, the estimates of screening test error rates (in particular the estimated false negative rate) are usually more accurate than those obtained when assuming that the confirmatory test is error free (see Column 3 in Table 2).

The single population design also yields quite robust estimates for the prevalence rate and reasonable estimates of the rate, c, of the relative error rates of the confirmatory test to those of the screening test. The bias in the prevalence rate estimates are less than .008 across the 10 examples presented and these estimates are considerably more accurate than the estimates from the uncorrected confirmatory test (Column 3). The estimate of the common reduction in error from the confirmatory test lies between the actual rates, although it is not the midpoint.

DISCUSSION

The Hui and Walter (2) method has proven to be an effective method for evaluating the error rates from diagnostic testing when its underlying assumptions are met. Our results show that the method yields a robust estimator of the prevalence when the two subpopulations have different error rates provided that the prevalence rates are quite different. In contrast, the estimates of the sensitivity and specificity of the screening test, and the confirmatory test to an even greater degree, were seen to be affected by modest violations of the assumption that the error rates in the two subpopulations were the same. The method appears to be less sensitive to this assumption when the error rates of both tests are larger in the subpopulation with the lower prevalence.

An alternative method that randomly splits a sample into two subgroups, but assumes that the confirmatory test reduces the error rates of the screening test by the same fraction (EFR), is proposed. The estimates of both the accuracy of the screening test and the prevalence rates were robust to moderate violations of the EFR assumption. Furthermore, the design provides an estimate of the relative error in the confirmatory that approximates the average reduction in the two error rates when the EFR assumption is violated.

When three tests are available, Irwig and Walter (10) showed that one can estimate the accuracy rates of the tests and the prevalence rates without assuming a relationship between the test error rates and without the

TABLE 3
Evaluation of the Single Population Method When the EFR Assumptions are Violated. Actual Screening False Positive Rate and a False Negative Rate are Equal to .05

EXAMPLE	True Prevalence Rate	Confirmatory Test Reduction in Error Rates		Uncorrected Confirmatory Test Estimate of			Single Population Method Estimates Standard Error of Estimates in ( )s				Bias in Single Population Estimate of π
		Ratio of α_2 to α_1	Ratio of β_2 to β_1	π	α_1	β_1	α_1	β_1	c	π	
1	.20	.25	.25	.20750	.05284	.09337	.05 (.00792)	.05 (.03205)	.25 (.16609)	.20 (.01348)	.00000
2	.20	.25	.350	.20675	.05341	.09348	.05123 (.00716)	.04504 (.03183)	.26989 (.16755)	.19781 (.01372)	.00219
3	.20	.350	.25	.21150	.05285	.10957	.04866 (.00796)	.05534 (.03502)	.32988 (.18262)	.20239 (.01504)	.00239
4	.20	.25	.500	.20500	.05566	.20652	.05320 (.00703)	.03697 (.03133)	.29929 (.15984)	.19432 (.01351)	.00567
5	.20	.500	.25	.21750	.05288	.13276	.04590 (.01103)	.06605 (.04304)	.44909 (.20793)	.20731 (.01948)	.00731
6	.20	.25	.15	.20850	.05171	.09317	.04883 (.00694)	.05465 (.03134)	.22990 (.16733)	.20208 (.01354)	.00208
7	.20	.25	.05	.20950	.05057	.09296	.04772 (.00700)	.05902 (.03108)	.20960 (.16617)	.20406 (.01381)	.00406
8	.02	.25	.350	.03190	.05033	.39561	.05012 (.00484)	.04407 (.11840)	.25199 (.10975)	.01974 (.03055)	.00026
9	.02	.25	.500	.03175	.05047	.39724	.05030 (.00471)	.03496 (.11965)	.25492 (.10787)	.01935 (.02750)	.00065
10	.02	.500	.25	.04425	.05023	.54831	.04955 (.00634)	.07131 (.15810)	.49488 (.22974)	.02098 (.04790)	.00098

need for a second subpopulation with a different prevalence. All of these methods assume that the test results are independent given the individual's true status. In light of the analysis conducted by Vacek (11) of the Hui and Walter method (2) when this assumption is violated, one expects that this conditional independence assumption is essential. A separate study should be conducted in conjunction with the designs discussed in this paper to evaluate whether this assumption is met. Other designs along the lines of Sinclair and Gastwirth (12) that estimate the dependence parameters could be used in this process.

Acknowledgments—This work is supported by a grant from the National Science Foundation. We wish to thank S. D. Walter and S. L. Hui for providing us with the technical details concerning the derivation of their method as well as sharing their experiences with the approach.

REFERENCES

1. Hui SL, Xiao HZ. Evaluation of diagnostic tests without a gold standard. *Stat Methods Med Res.* 1998;7:354–370.
2. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics.* 1980;36:167–171.
3. Goldberg JD. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *J Am Stat Assoc.* 1975;70:561–567.
4. Gastwirth JL. The statistical precision of medical screening procedures: application to polygraph and aids antibodies test data. *Stat Science.* 1987;2(3): 213–238.
5. SAS Institute Inc. *SAS/Stat User's Guide, release 6.03 edition.* Cary, NC: SAS Institute, Inc; 1988.
6. Bard Y. *Nonlinear Parameter Estimation.* New York: Academic Press; 1974.
7. Jennrich RI, Moore RH. Maximum likelihood estimation by means of nonlinear least squares. *American Statistical Association: Proceedings of the statistical computing section.* 1975;57–65.
8. Bradley EL. The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. *J Am Stat Assoc.* 1973;68:199–200.
9. Sinclair MD. *Measuring response errors in surveys.* Unpublished doctoral dissertation. Washington DC: George Washington University; 1994.
10. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence, and relative risk from misclassified data: A review. *J Clinical Epidemiol.* 1988;41: 923–937.
11. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics.* 1985; 41:959–968.
12. Sinclair MD, Gastwirth JL. On procedures for evaluating the effectiveness of reinterview survey methods: Application to labor force data. *J Am Stat Assoc.* 1996;91(435):961–969.